Pipeline Pilot Interface

Version 1.0.0.4

User Guide

(for Package version 1.0.0 and above and Pipeline Pilot version 8.0 and above)

Edgar Derksen, Sally Hindle

© 2015 BioSolveIT GmbH, An der Ziegelei 79, 53757 St. Augustin, Germany
Phone ++49-2241-2525-0, support@biosolveit.de

Contents

Contents **3**

1 Quick Start Steps **5**

1.1 Download and Import current PipelinePilot Package 5

1.2 Install Tools 5

1.3 Create or Update Protocols 5

2 Installation **7**

2.1 BioSolveIT Web Installation 7

2.2 Using Offline Installation Components 7

2.3 Using an External Tool Installation 7

2.3.1 To Connect to an Tools Installation on the Pipeline Pilot Server 8

2.3.2 To Connect to Tools Installation on a Remote Linux Server 9

2.4 Running Tools in Parallel in Pipeline Pilot 10

2.4.1 Running in Parallel on the Pipeline Pilot Server 11

2.4.2 Running in Parallel on a Remote Linux Cluster 11

2.4.3 Example Scenarios and Required Settings 14

2.5 Uninstallation 15

3 Trouble Shooting **17**

3.1 Problems using the *ssh* Method 18

3.2 Further help and BioSolveIT PDF Reporter 18

4 Tips and Tricks **21**

4.1 Other Significant Parameters in at Components 21

4.1.1 <Has Same File System> 21

4.2 Accessing Other Domains within Pipeline Pilot 21

1

Quick Start Steps

1.1 Download and Import current PipelinePilot Package

1. Download current PipelinePilot package on

   http://www.biosolveit.de/download/
2. Extract PipelinePilot package to a custom directory.
3. From the custom directory import all components (.xml) using PP-Client via drag and drop to your components collection.

1.2 Install Tools

1. Drag and drop BioSolveIT Web Installer component from the components collection to an empty protocol.
2. Set up Parameter "Client Destination Dir". Use the "..." facility, within filebrowser popup click on dektop icon and press "select" at the bottom right. Using this setting the components will be downloaded to your local desktop.
3. Add your BioSolveIT license in the field for the parameter <Run Tool -> on PP Server -> License Server or License File> in the Implementation tab – if your license is available from a license server, simply type in the name of the server in this format *@server-name*, or if you have a license file then you may browse for it using the ... facility.
4. Run protocol once, it will result into html reports showing component download and installation success.
5. For alternative installations and more details see chapter 2.

1.3 Create or Update Protocols

1. Import components now existing in the defined "Client Destination Dir" via drag and drop to your components collection.
2. Drag and drop new components into a new protocol.

3. Update via a *right click drag and drop* of new component onto component reference in a protocol.
4. Now you are ready to setup the new components and run your protocol.

2

Installation

2.1 BioSolveIT Web Installation

The easiest way to download and install BioSolveIT packages and tools is to download and run BioSolveIT Web Installer component. Download BioSolveIT in PipelinePilot package from http://www.biosolveit.de/download. Read UserGuide within that package for further details.

Note that both PP-Client and PP-Server need web access to run BioSolveIT Web Installer. Without web access you need to follow one of the following sections in this chapter.

2.2 Using Offline Installation Components

For an offline installation working similar to BioSolveIT Web Installation please download for example: LeadIT in PipelinePilot package and use the contained LeadIT Installer component. For further details read also the Userguide of LeadIT in PipelinePilot package.

2.3 Using an External Tool Installation

You can also use an existing external installation of the Tool software at each component. This means you also have the opportunity to use settings different to those set by default. To do this, you must already have Tools installed somewhere on your system outside of Pipeline Pilot. To install Tools yourself, visit the download page at BioSolveIT:
`http://www.biosolveit.de/download`
and fetch the download package for your system for the latest Tools package. Follow the instructions in the package to install Tools and receive your licenses. Enter the license information for Tools as described in the package **and not using the parameter <Run Tool -> on PP Server -> License Server or License File> as for the internal installation**.

To use an external installation of Tools, you must change the value of the parameter <Run Tool-> on PP Server -> Use> in the Implementation tab to *preinstalled Tool*.

There are actually two ways to use Tools with an external installation. These are by using Tools installed directly on the Pipeline Pilot server , or by accessing a remote machine where Tools are installed using *ssh* method. The method is selected using the parameter <Run Tool> Both methods are covered in more detail below.

2.3.1 To Connect to an Tools Installation on the Pipeline Pilot Server

The most common scenario is that you will have an installation of Tools on the Pipeline Pilot server. If you choose this option, you just have to enter the path to the executable and configuration file as parameters in the Implementation tab. Pipeline Pilot will then just start Tools whenever they are required by making a call to the executable that you entered.

This method is given the name *on PP Server* – On Pipelinepilot Server. You can see an example in figure 2.2.

- Requirement:
  - Tools are installed on the Pipeline Pilot server. You can see which machine is the Pipeline Pilot server by starting your copy of Pipeline Pilot Client on your own workstation and find the name or IP of the server shown at the bottom right of the status bar (see figure 2.1). You must find where the tool installations are *on that machine*.

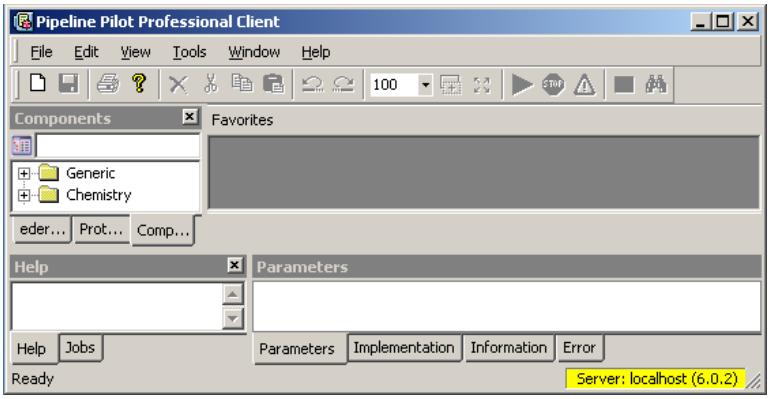

Figure 2.1: See where your Pipeline Pilot Server is installed.

- General Steps:
  1. Set <Run Tool> to *on PP Server*
  2. Set <Run Tool-> Use> to *Preinstalled Tool*
  3. For the parameter <Run Tool-> on PP Server -> Executable>, enter the path to the tool installation on the Pipeline Pilot server. For example: `C:\Program Files\BioSolveIT\tools3\tools.exe`
  4. Add your license for the executable of Tools in the field for the parameter <Run Tool-> on PP Server -> License Server or File> in the Implementation tab – if your

license is available from a license server, simply type in the name of the server in this format *@servername*, or if you have a license file then you may browse for it using the ... facility.

You can save these settings in the component.

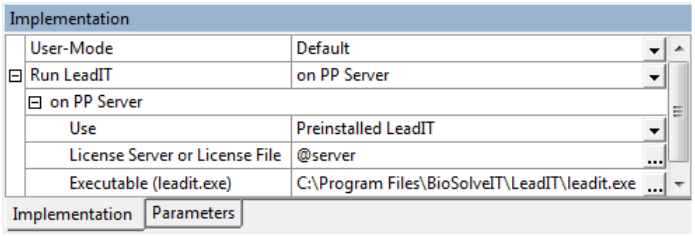

Figure 2.2: *PPServer*: Connect to Tools Installation on the Pipeline Pilot Server

2.3.2 To Connect to Tools Installation on a Remote Linux Server

Your existing Tools installation could be on a Linux computer remote from the Pipeline Pilot server – in this case we offer an alternative so you can use the remote installation instead. Here, the calculations carried out by Tools will be done on the remote Linux machine. Pipeline Pilot logs into the Linux machine using `ssh`, having copied all relevant files to the machine, and will run the calculation there, finally copying back all data it needs to the Pipeline Pilot server to continue with the pipeline.

This method is given the label *ssh*. You can see an example in figure 2.3.

- Requirement:
  - Tools are installed on a Linux machine available to the Pipeline Pilot server via `ssh`.
- General Steps:
  1. Set <Run Tool> to *via SSH*
  2. For the parameter <Run Tool-> via SSH -> Executable>, enter the path to the Tools installation on the Linux machine. For example: `/software/BioSolveIT/tools/bin/tools`
  3. For the parameter <sshHost>, enter the Linux machine host name
- User specific steps:
  1. For the parameter <Run Tool-> via SSH -> User>, enter the user login name for ssh on the Linux machine
  2. For the parameter <Run Tool-> via SSH -> Password>, enter the user password for ssh on the Linux machine

3. There are more advanced options to be found under <Run Tool-> via SSH -> Options> for more specific ssh parameters. Note the option <Run Tool-> via SSH -> Options -> Delete Results>, this may be useful for trouble-shooting later.

You can save these settings in the component (be sure not to save your own user specific login details in components available to others!).

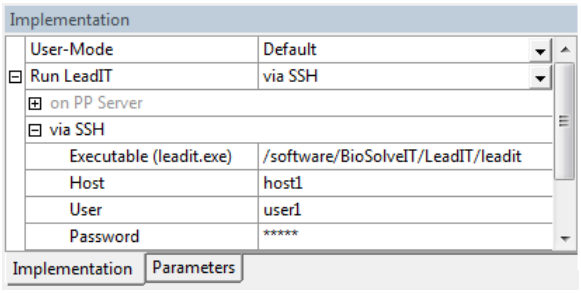

Figure 2.3: *ssh*: Connect to Tools Installation on a Remote Linux Server via ssh

All files necessary for the calculation will be transferred via `scp` between the Pipeline Pilot server and ssh Linux machine. Files copied and files created the remote server are automatically deleted at the end of the job leaving no trace. However, in case the user would like to leave a copy of the calculation and result files on the Linux machine, or for trouble-shooting as mentioned above, it is possible to set a parameter to tell Pipeline Pilot not to delete these files:
<Run Tool-> via SSH -> Options -> Delete Results> : *False*

2.4 Running Tools in Parallel in Pipeline Pilot

Tools in Pipeline Pilot takes advantage of the parallel computing options available in Pipeline Pilot to speed up longer calculations. This section tells you how to adjust the options to your system and needs. You will find the options in the Implementation tab, as in figure 2.4.

The most important limitation to a parallel processing calculation is the number of licenses that you have. If you only have a single license then parallel calculations will not be possible. Further choices in the set-up of the parallel computing computing calculation depend on the number of Pipeline Pilot licenses you have plus your choice of connection to the external Tools installation (*on PP Server* or *via SSH*).

It is important to note that a balance must be achieved between the overhead caused by running several calculations instead of one and the size of the calculation – there is a lot of overhead involved in sending all the essential data to different computers and collecting the results. For the Tools component, however, we still think parallel computing will work also for parallel jobs with batch size 1. Of course, these figures depend on the speed of your

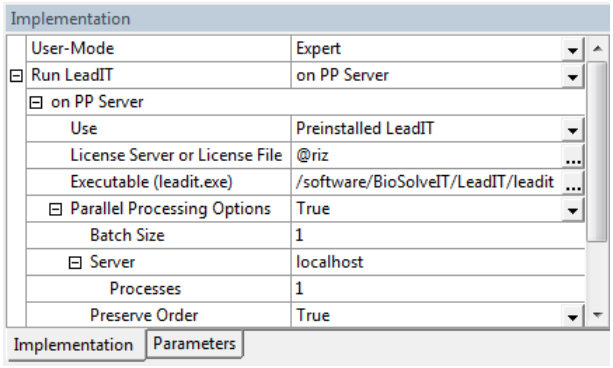

Figure 2.4: The options for tuning parallel processing are found in the Implementation tab

machines and network: experiment with your set-up if you intend to carry out large calculations often. Also we advise you to read the Pipeline Pilot documentation about parallel processing to understand more fully how it works.

Note: to set up large parallel processing jobs, you will need Administrator rights to change one setting.

2.4.1 Running in Parallel on the Pipeline Pilot Server

You may have a multi-processor machine as your Pipeline Pilot server. If you also have the appropriate number of Pipeline Pilot and Tools licenses the simplest way to start a parallel calculation is to raise the number of processes to the number of processors of the machine.

You could also have more than one Pipeline Pilot server available in your network. If so, you can enter a list of the server names at the parameter <Server>. Below that, for the parameter <Processes> enter a list of the number of processes each server should receive. The lists are both comma separated and must be in corresponding order. Remember to adjust also the <Batch Size> accordingly.

Note: the path to the external installation of Tools must be same on all servers!

2.4.2 Running in Parallel on a Remote Linux Cluster

We have developed an implementation in the component whereby a large cluster can be incorporated to run tool jobs, without them having to be Pipeline Pilot servers. However, it must be a Linux cluster and the component must use the ssh method. You must also have enough tool licenses available to the cluster.

The settings in the Implementation tab must be made as for running the ssh method, with two important changes. Instead of one <Run FTrees -> via SSH -> Host>, enter a comma separated list of the host names in the Linux cluster. Then, you must choose how many batchs in total you want the job to be split into and enter this total in the <Processes> parameter. Remember to change the <Batch Size> to fit with the number of processes.

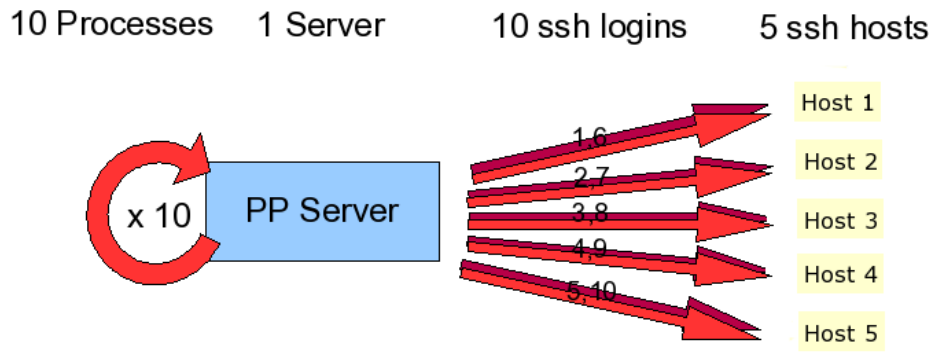

Figure 2.5: The method of parallel processing on Linux clusters in the tool component. The Pipeline Pilot server is given 10 processes. The 10 processes start an ssh job respectively, distributed amongst the ssh hosts.

Figure 2.5 may help clarify how the method works.

The Pipeline Pilot server will run with the number of processes given in <Processes>. However, in this case the processes the server receives are not the tool calculations themselves but instructions for starting the ssh jobs. The server iterates through its 10 jobs each time spawning an ssh job on a Linux host. Beware that the Pipeline Pilot server does not know how many processors the Linux hosts have so make sure you choose the number of <Processes> to fit the number of Linux hosts and their number of processors respectively – be careful not to overload the Linux hosts.

Figure 2.6 shows the detailed settings required to set up the calculation as in figure 2.5.

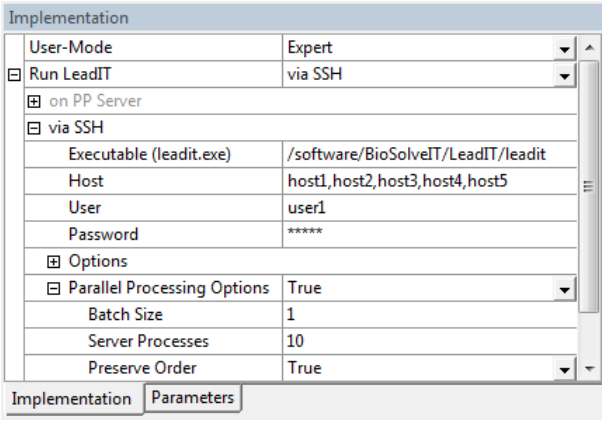

Figure 2.6: The options for parallel processing for the example shown in figure 2.5 would look similar to these

As you may already have realised, you could enter more than one Pipeline Pilot server at the <Server> parameter (along with another entry for number of <Processes> as a comma

separated list) to execute a doubly parallel calculation!

Note: to get this method to work you will need to change the maximum number of processes per Pipeline Pilot server. The Pipeline Pilot Client will let you enter any number for the <Processes> parameter and does not warn you if this number is above the maximum.

Changing the maximum number requires Administrator rights! The number of processes per Pipeline Pilot server is usually restricted to the number of processors of the server. You must override this maximum to be able to set the number of processes you want for your parallel calculation. In the above example, the maximum must be set to 10 or more. Take the following steps:

- Go to the 'Scitegic Server Home Page', for example, via the Help menu in your Pipeline Pilot client.
- Click on 'Pipeline Pilot Administration Portal' and log in with the Administrator user name and password.
- In the last field of the table ('Maximum number of simultaneous parallel processing subprotocols allowed') change the value, click 'Save' and log out again.

2.4.3 Example Scenarios and Required Settings

ssh/PPServer	*PPServer*
Parallel Processing Options	False
Number of PP servers in list	1
Number of ssh Hosts in list	-
Number of Processes	-
Behavior	The calculation will run as one complete job on the Pipeline Pilot server
ssh/PPServer	*ssh*
Parallel Processing Options	False
Number of PP servers in list	-
Number of ssh Hosts in list	1
Number of Processes	-
Behavior	The calculation will run as one complete job on the ssh host
ssh/PPServer	*PPServer*
Parallel Processing Options	True
Number of PP servers in list	1
Number of ssh Hosts in list	-
Number of Processes	1
Behavior	(Default) Induces the 'pipeline' effect: the job will run in chunks on the server processor(s)
ssh/PPServer	*PPServer*
Parallel Processing Options	True
Number of PP servers in list	1
Number of ssh Hosts in list	-
Number of Processes	>1
Behavior	A true parallel effect: the job will be run in chunks on the server processors
ssh/PPServer	*PPServer*
Parallel Processing Options	True
Number of PP servers in list	4 (each with two processors)
Number of ssh Hosts in list	-
Number of Processes	2,2,2,2
Behavior	A true parallel effect: the job will run in chunks in parallel split across 8 processors
ssh/PPServer	*ssh*
Parallel Processing Options	True
Number of PP servers in list	1
Number of ssh Hosts in list	5 (each with 2 processors)
Number of Processes	10
Behavior	A true parallel effect: the job will run in chunks in parallel split across 10 processors

2.5 Uninstallation

Use BioSolveIT Uninstaller component to uninstall all Tools.

Note that this kind of deinstallation only removes Tools installed by installer component or BioSolveIT web installer component.

3

Trouble Shooting

The most commonly seen problems with Tools in Pipeline Pilot is with the connection to the external Tools installation. For one thing, Tools itself must be correctly installed on the system independently of Pipeline Pilot – it is essential first to make sure this is the case (especially to make sure that Tools can locate the licenses). Once Tools runs fine on your system, the remaining key task is to make sure the path to the executable is correct within the Pipeline Pilot component.

When the error messages pop up, they may contain an ToolError in the error message box, as in figure 3.1.

Go to the 'Jobs' tab below the Protocol workspace and check under the last run job for a file called 'ToolComponent Debug' – as in figure 3.2. Clicking on the link brings up HTML report with input and output data in a browser.

A correctly started Tools job outputs the following header: if there is a problem you will see some of this header and the point where the problem occurs:

```
                                L e a d I T
Copyright               Prediction of Protein-Ligand Interactions

BioSolveIT GmbH         Version:   1.0.0.1
An der Ziegelei 79      Modules:   [CDOCK] [FLEXE] [PHARM] [SCREEN] [PERMUTE]
53757 St. Augustin
Germany                 Original Author:    Matthias Rarey
www.biosolveit.de       Contact:            leadit@biosolveit.de

 For information about additional contributors and copyright notes
 please consult the user guide or type 'help about'.

 >> Running on DELTA (Windows 5.1) with 2 processors.
 >> Loaded settings (v 1.1.0) from 'C:\userdir\.leadit\settings.pxx'.
 >> LeadIT_base license check (BioSolveIT keys): succeeded.
 >> Licensed modules: LeadIT [CDOCK] [FlexE] [PHARM] [PERMUTE] [SCREEN] [DEVELOP]
...
```

3.1 Problems using the *ssh* Method

You may also experience problems using the *ssh* login, for example, the user name is unknown or the host is not found.

3.2 Further help and BioSolveIT PDF Reporter

More complicated errors may arise during the running of tool. Again though, the errors will be collected and as much information shown as possible. If you are familiar with the tool you may want to take a look at all the output of the job yourself to see if you can recognize the problem. In this case, you can look in the temporary folders Pipeline Pilot sets up internally to find the output, or if you are working with the *ssh* method, set the parameter <Run FTrees -> via SSH -> Options -> Delete Results> to *False* so that you may then find the files retained on the *ssh* host: these will be in the directory set under the ssh parameter <Run FTrees -> via SSH -> Options -> Temp Path> (see the help text associated with this parameter to find its default value) – essentially a cryptically named folder whose name begins with the date and time of the job!

If you still do not know what is causing the errors, write down as much information as possible relating to your installation scheme. You can also create a PDF Report unsing BioSolveIT PDF Reporter component which summarizes installation data into one pdf file. Send all the information to:

support@biosolveit.de

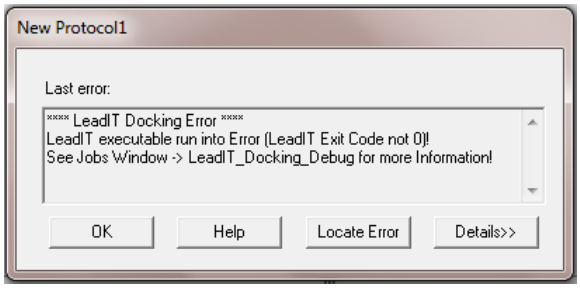

Figure 3.1: An error box reporting that the Tools exited with error.

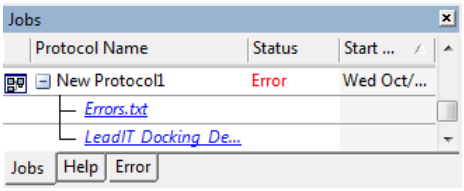

Figure 3.2: The full error report can be found in the 'Jobs' tab.

4

Tips and Tricks

4.1 Other Significant Parameters in at Components

For detailed documentation of all parameters, refer to the documentation you find in the *Help* area of the Pipeline Pilot window.

We list here particularly interesting parameters: those that greatly influence the workflow or change the outcome of calculations, or those that may help you understand what is happening in the component.

4.1.1 <Has Same File System>

Implementation : <Run Tool-> via SSH -> Options -> Has Same File System>

Normally, for an *ssh* job, Pipeline Pilot must first copy all the data required by the tool to the <Run FTrees -> via SSH -> Host> using scp. This is time consuming. It is possible that the Pipeline Pilot Server and <Run FTrees -> via SSH -> Host> actually share the same file system rendering the scp process unnecessary. Select True if the Pipeline Pilot Server and <Run FTrees -> via SSH -> Host> share the same File System - no copying of data is necessary. Selecting False means Pipeline Pilot copies all the data to and back from the <Run FTrees -> via SSH -> Host> – this is just a little slower but will always still work. Leave the parameter set to False if you are uncertain!

4.2 Accessing Other Domains within Pipeline Pilot

Often in house data or even your own working data are accessible from a windows computer via a domain (a path starting for example 'z:\...' or '\\...') which you cannot find from within Pipeline Pilot. That means you must first literally transfer the data to the Pipeline Pilot Server itself.

If you are using a Linux Pipeline Pilot server, this hint does not apply.

To get around this problem and make the Pipeline Pilot working environment much more flexible you can allow users access to domains – you need Administrator rights to be able to do this! Also check first that you should change these settings as they may have already been set to fit the current environment.

- Go to the 'Scitegic Server Home Page', for example, via the Help menu in your Pipeline Pilot client.
- Click on 'Pipeline Pilot Administration Portal' and log in with the Administrator user name and password.
- In the Security tab go to Authentification.
- For the 'Authentification Method' choose 'DOMAIN' and a set of parameters will appear.
- Enter the domain name in the field 'Domain' and choose 'Full' for 'Impersonation'
- Choose 'DOMAIN' for 'Retrieve Groups' and leave 'Limit access to listed domains' set to 'No'

- click 'Save' and log out again.

After you have done this you will need to enter your domain login details when you start the Pipeline Pilot Client.